Journal of Neuroscience Methods 235 (2014) 277–284

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Computational Neuroscience

SCOPRISM: A new algorithm for automatic sleep scoring in mice

Stefano Bastianini[a], Chiara Berteotti[a], Alessandro Gabrielli[c], Flavia Del Vecchio[b], Roberto Amici[b], Chloe Alexandre[d], Thomas E. Scammell[d], Mary Gazea[e], Mayumi Kimura[e], Viviana Lo Martire[a], Alessandro Silvani[a], Giovanna Zoccoli[a,*]

[a] PRISM Lab, Alma Mater Studiorum – University of Bologna, Bologna, Italy
[b] Physiological Regulation in Wake-Sleep Cycle Lab, Department of Biomedical and Neuromotor Sciences, Alma Mater Studiorum – University of Bologna, Bologna, Italy
[c] Department of Physics and Astronomy, Alma Mater Studiorum – University of Bologna, Bologna, Italy
[d] Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, USA
[e] Max Planck Institute of Psychiatry, Munich, Germany

HIGHLIGHTS

- We tested a new open-source sleep-scoring algorithm (SCOPRISM) on 92 mice.
- We successfully validated SCOPRISM in wild-type mice and mouse models of obesity and narcolepsy.
- We cross-validated SCOPRISM on mice and rats recorded and analyzed in other labs.
- We developed a quick and easy visual flow-chart for the correct use of SCOPRISM.

ARTICLE INFO

Article history:
Received 4 June 2014
Received in revised form 23 July 2014
Accepted 24 July 2014
Available online 1 August 2014

Keywords:
Sleep
Scoring
Validation
Mice
Rat
Narcolepsy

ABSTRACT

Background: Scoring of wake–sleep states by trained investigators is a time-consuming task in many sleep experiments. We aimed to validate SCOPRISM, a new open-source algorithm for sleep scoring based on automatic graphical clustering of epoch distribution.
Methods: We recorded sleep and blood pressure signals of 36 orexin-deficient, 7 leptin knock-out, and 43 wild-type control mice in the PRISM laboratory. Additional groups of mice ($n = 14$) and rats ($n = 6$) recorded in independent labs were used to validate the algorithm across laboratories.
Results: The overall accuracy, specificity and sensitivity values of SCOPRISM (97%, 95%, and 94%, respectively) on PRISM lab data were similar to those calculated between human scorers (98%, 98%, and 94%, respectively). Using SCOPRISM, we replicated the main sleep and sleep-dependent cardiovascular findings of our previous studies. Finally, the cross-laboratory analyses showed that the SCOPRISM algorithm performed well on mouse and rat data.
Comparison with existing methods: SCOPRISM performed similarly or even better than recently reported algorithms. SCOPRISM is a very simple algorithm, extensively (cross)validated and with the possibility to evaluate its efficacy following a quick and easy visual flow chart.
Conclusions: We validated SCOPRISM, a new, automated and open-source algorithm for sleep scoring on a large population of mice, including different mutant strains and on subgroups of mice and rats recorded by independent labs. This algorithm should help accelerate basic research on sleep and integrative physiology in rodents.

Abbreviations: BP, blood pressure; EEG, electroencephalographic; EMG, electromyographic; HTG, orexin-ataxin3 narcoleptic transgenic mice with hybrid genetic background; KO, orexin knock-out narcoleptic mice; NREMS, non-rapid-eye movement sleep; Ob/ob, leptin knock-out obese mice; REMS, rapid-eye movement sleep; TG, orexin-ataxin3 narcoleptic transgenic mice with pure genetic background; W, wakefulness; WT, wild-type mice.

* Corresponding author at: Dipartimento di Scienze Biomediche e Neuromotorie, Alma Mater Studiorum – Università di Bologna, Piazza di Porta San Donato 2, 40126 Bologna, Italy. Tel.: +39 051 2091726; fax: +39 051 2091737.
E-mail address: giovanna.zoccoli@unibo.it (G. Zoccoli).

http://dx.doi.org/10.1016/j.jneumeth.2014.07.018
0165-0270/

1. Introduction

Scoring wake–sleep states based on electroencephalographic (EEG) and electromyographic (EMG) recordings is a time-consuming process, yet reliable scoring is critical in sleep research. Pre-clinical sleep laboratories increasingly investigate mouse models because of the power of genetic tools applicable to this species. Most of these laboratories base their studies on manual sleep scoring by trained investigators (El Helou et al., 2013; Gondard et al., 2013; Kantor et al., 2013). However, this often becomes the experimental bottleneck and a potential source of subjectiveness affecting research outcomes. To overcome these difficulties, different commercial or open-source algorithms for automatic sleep scoring have been proposed in the last few years (Brankack et al., 2010; Rytkonen et al., 2011; Sunagawa et al., 2013; Veasey et al., 2000). Some of these algorithms are computationally intensive (Sunagawa et al., 2013), or have not been tested on independent datasets (i.e. cross-laboratory validation) (Brankack et al., 2010; Rytkonen et al., 2011; Sunagawa et al., 2013; Veasey et al., 2000). Subjectiveness still represents a problem for those algorithms that require a pre-stage of manual scoring on a subset of recording data (Rytkonen et al., 2011). Finally, validation of these algorithms has been performed either on a limited number of mice ($n = 6$–9) (Brankack et al., 2010; Rytkonen et al., 2011; Sunagawa et al., 2013) or only on non-mutant mice only (Brankack et al., 2010; Rytkonen et al., 2011; Veasey et al., 2000). Importantly, a sleep-scoring algorithm have not been validated in mouse models of narcolepsy (Chemelli et al., 1999; Hara et al., 2001), one of the most intensively studied human sleep disorders. On the other hand, recent technical tools, such as telemetric devices, allow researchers to measure cardiovascular and respiratory variables simultaneously with EEG and EMG (Bastianini et al., 2011; Lo Martire et al., 2012; Silvani et al., 2009). The development of an automatic sleep-scoring algorithm tested on mutant mice with multiple physiological recordings would thus accelerate integrative physiology as well as behavioral studies.

In the past few years, we have investigated sleep structure and sleep-dependent cardiovascular control in different strains of mutant mice, including leptin-deficient mice with genetic obesity and hypocretin (orexin) deficient mice as a model of narcolepsy (Bastianini et al., 2011; Lo Martire et al., 2012; Silvani et al., 2009). All these studies involved manual sleep scoring at a high (4 s) temporal resolution by trained investigators. To speed up the procedure, manual sleep scoring consisted of using raw EEG and EMG recordings to correct or confirm a suggestion provided by an automatic sleep-scoring algorithm, which we had developed for this purpose.

Here, we aimed to retrospectively validate the performance of our automatic sleep-scoring algorithm (SCOPRISM) in correctly discriminating wake–sleep states in different mouse strains. In particular, our validation procedure consisted of 3 main steps: (a) we evaluated the algorithm performance in terms of accuracy, specificity, and sensitivity of wake–sleep state discrimination. We paid particular attention to algorithm sensitivity, which is usually a point of weakness of automatic algorithms (Rytkonen et al., 2011; Sunagawa et al., 2013), especially in terms of rapid-eye-movement sleep (REMS) discrimination. (b) We tested whether differences between groups of mutant mice, which we previously found and published in terms of sleep structure and sleep-dependent cardiovascular control employing manual sleep scoring (Bastianini et al., 2011; Lo Martire et al., 2012; Silvani et al., 2009), would still have been significant had we relied on automatic sleep scoring only. (c) We performed a cross-laboratory validation evaluating the robustness of the SCOPRISM algorithm on mouse and rat data recorded and analyzed by independent research teams.

2. Materials and methods

The study protocols were approved by the Bologna University ethics committee on animal experimentation and complied with the National Institutes of Health guide for the care and use of laboratory animals.

2.1. Mouse strains

The study involved a total of 86 male mice recorded in the PRISM lab and already described in other publications (Bastianini et al., 2011; Lo Martire et al., 2012; Silvani et al., 2009). In particular, groups consisted of: (a) hypocretin-ataxin3 transgenic (TG) narcoleptic mice (Hara et al., 2001) with genetic ablation of hypocretin neurons and with pure (TG, $n = 12$) or hybrid (Hara et al., 2005) (HTG, $n = 16$) C57BL/6J genetic background; (b) hypocretin gene knock-out mice (Chemelli et al., 1999) (KO, $n = 8$); (c) leptin-deficient mice (ob/ob, $n = 7$, Harlan Laboratories, Holland); (d) merged group of all WT controls ($n = 43$) including mice with pure ($n = 26$) or hybrid ($n = 17$) C57BL/6 genetic background. For cross-lab SCOPRISM validation, we also analyzed two additional groups of WT mice recorded and scored at Beth Israel Deaconess Medical Center (T.E.S. and C.A; $n = 8$; mice were purchased from Jackson Lab, Bar Harbor, ME, USA) and at the Max Planck Institute of Psychiatry (M.K. and M.G.; $n = 6$; C57BL/6N mice were bred in the facility of Max Planck Institute of Biochemistry, Martinsried, Germany). Finally, for cross-species validation of SCOPRISM, we analyzed a group of WT (Sprague-Dawley) rats recorded and analyzed in the lab of physiological regulation in wake–sleep cycle, Department of Biomedical and Neuromotor Sciences in Bologna, Italy (R.A. and F.D.V; $n = 6$; rats were purchased from Charles River Italy).

2.2. Surgery, sleep recordings, and data acquisition

All the mice recorded in the PRISM lab were implanted with a pair of stainless-steel miniature screws (00-96 × 3/32, Plastics One, Roanoke, 6 VA, USA) put in contact with the dura mater to obtain an ipsilateral fronto-parietal EEG signal (differential derivation). The frontal screw was placed 1 mm anterior and 1 mm lateral to bregma. The parietal screw was placed 1 mm anterior and 1 mm lateral to lambda. A pair of Teflon-coated stainless steel wires (Cooner Wire, Chatsworth, CA, USA) was inserted in the posterior neck muscles to record the EMG signal. The EEG and EMG signals were transmitted with a cable connected to a rotating swivel (SL2 + 2C/SB, Plastics One, Roanoke, 6 VA, USA). Mice were also implanted with a telemetric blood pressure (BP) transducer (TA11PA-C10, DSI, Tilburg, the Netherlands) connected to a catheter inserted into the abdominal aorta. Simultaneous recordings of EEG, EMG and BP were performed for at least 44 h on mice freely behaving in their own cages. Ambient temperature during recordings was always set at 25 °C except for 10 HTG and 8 WT mice, which were recorded at 30 °C (Lo Martire et al., 2012). The EEG and EMG signals were amplified, filtered (EEG: 0.3–100 Hz; EMG: 100–1000 Hz; 7P511J amplifiers, Grass, West Warwick, RI, USA), sampled at 1024 Hz, and down-sampled at 128 Hz for data storage. The EEG and EMG amplifier gains were adjusted for each mouse to avoid signal saturation. The BP signal was sampled at 1024 Hz. For mouse recordings in the T.E.S. lab, epidural stainless steel screws electrodes (Plastic Ones) were implanted for ipsilateral frontoparietal EEG recordings (differential derivation; 1.5 mm lateral and 1 mm anterior to bregma; 1.5 mm lateral and 1 mm anterior to lambda). EMG electrodes were made from fine, multi-stranded stainless steel wire (AS131; Cooner Wire), and were inserted into the neck extensor muscles. All electrodes were attached to a micro-strip connector affixed to the animal's head with dental cement. EEG/EMG signals were acquired

using Grass Instruments model 12 amplifiers (×5000; filtered at 0.3–100 Hz) and digitized at 128 Hz. Signals were then digitally filtered (EEG, 0.3–30 Hz; EMG, 20–100 Hz) using Sleep-Sign (Kissei Comtec). Mice were continuously recorded for 24 h. For mouse recordings in the M.K. lab, ipsilateral frontoparietal EEG recordings (differential derivation) were obtained with epidural implantation of a frontal gold-wire electrode 1.5 mm lateral and 1.5 mm anterior to bregma and a parietal gold-wire electrode 3 mm lateral and 1 mm anterior to lambda. The surgical details followed those previously described (Fenzl et al., 2007; Kimura et al., 2010). The EEG and EMG signals were amplified 10,000×, filtered (EEG, 0.25–64 Hz; EMG, 175–1000 Hz) and digitized at 128 Hz using a LabVIEW program (National Instruments, Austin, TX) especially designed for sleep EEG/EMG acquisition system (EGEraVigilanz, SEA, Cologne, Germany). The sleep data analyzed here were recorded for two separate baseline days (24 h each). For the rat recordings in the R.A. lab, ipsilateral frontoparietal EEG recordings (differential derivation) were obtained with epidural implantation of a frontal electrode 2 mm lateral and 2 mm anterior to bregma and a parietal electrode 2 mm lateral and 4 mm posterior to bregma. The surgical details followed those previously described (Cerri et al., 2013). The EEG and EMG signals were amplified (10× and 6×, respectively), filtered (EEG: 0.3–30 Hz,; EMG 100–1000 Hz), and digitalized (EEG: 500 Hz; EMG: 1000 Hz). The sleep data analyzed here were recorded for 48 h (2 separate baseline days, 24 h each).

2.3. Manual sleep scoring

For all mice recorded in the PRISM lab, manual scoring of wake–sleep states was performed by 3 trained investigators (S.B., C.B., V.L.M.) on all consecutive 4 s epochs. Manual scoring consisted of using raw EEG and EMG recordings to correct or confirm a suggestion provided by the automatic sleep-scoring algorithm. Wakefulness (W) was scored when the EMG tone was high and the EEG was at a low voltage with possible δ (0.5–4 Hz) and θ (6–9 Hz) frequency components. Non-rapid-eye-movement sleep (NREMS) was scored when the EMG tone was lower than in W and the EEG was at a high voltage with prominent δ frequency components. REMS was scored when the EMG indicated muscle atonia with occasional muscle twitches and the EEG was at a low voltage with predominant θ frequency components. Epochs with signals that were on a borderline between two different states were scored as "undetermined". According to a consensus definition (Scammell et al., 2009), a "cataplexy-like state" (CLE) was scored when an abrupt episode of nuchal atonia with predominant EEG θ rhythm lasted more than 10 s and was preceded by at least 40 s of active W. Sleep in each mouse was scored by a single investigator with the exception of 7 randomly-chosen mice of different strains (3 WT, 2 TG, and 2 ob/ob), which were scored separately by each investigator to assess the agreement between human scorers. Manual scoring in the T.E.S., M.K. and R.A. labs was performed on 10 s, 4 s and 1 s epochs, respectively. In the T.E.S. lab, manual scoring was performed on data pre-scored by another automatic algorithm (SleepSign).

2.4. Automatic sleep scoring

SCOPRISM, our automatic sleep-scoring algorithm, operated in 2 main steps with a time resolution of 4 s. In step 1, sleep scoring was drafted, according to two local properties of each 4 s epoch: the ratio between EEG spectral power in the θ (6–9 Hz) and δ (0.5–4 Hz) frequency ranges, and the root mean square (rms) of the EMG signal. In step 2, the sleep scoring of each epoch was refined, following the results of the scoring draft in adjacent epochs. The fraction of epochs characterized by each given combination of EEG θ/δ ratio and EMG rms was plotted in step 1 as 3D surface (Fig. 1, panels A and B). The 3D plot typically included a broad peak at high values of EMG rms, which corresponds to episodes of W, and a second sharper peak at low values of EMG rms and EEG θ/δ ratio, which corresponds to episodes of NREMS. This second peak typically declines progressively forming a ridge toward high values of the EEG θ/δ ratio, which corresponds to REMS. The automatic scoring draft (step 1) discriminated between W and sleep using only the EMG rms signal as an indicator. The boundaries of the valley between the W peak and NREMS peak on the 3D plot were set automatically by the algorithm (Fig. 1, panel C), and epochs with EMG rms values included in this boundary zone were assigned an indeterminate state. On the other hand, the discrimination between NREMS and REMS performed using the scoring draft (step 1) was based on values of the EEG θ/δ ratio <0.75 or >1.25, respectively. Epochs with low EMG rms values and with an EEG θ/δ ratio around unity (i.e., in a cutoff region between values of 0.75 and 1.25) were assigned an indeterminate state. The average amount of epochs scored belonging to the indeterminate state was less than 3.5% of the recording time. This step-1 scoring draft was refined by means of a set of simple rules

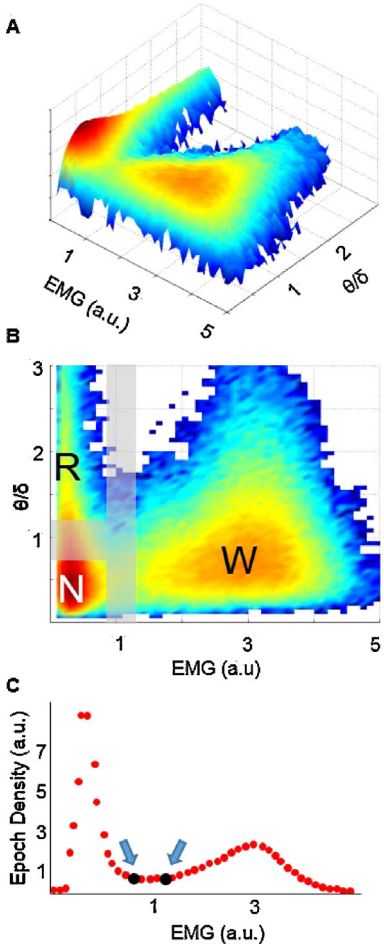

Fig. 1. Panel A shows a typical distribution profile of recorded 4 s epochs according to their electromyogram (EMG) root mean square values (x axis) and on the ratio between θ (6–9 Hz) and δ (0.5–4 Hz) spectral power of the electroencephalogram (EEG, y axis). On the z axis, a colorimetric scale represents the fraction of epoch characterized by each given combination of EMG and EEG values. For the sake of clarity, the y axis scale was truncated at $\theta/\delta = 3$. Panel B shows a bidimensional view of the same epoch distribution. Gray areas correspond to boundary regions (undetermined state attribution), which are used by the algorithm to discriminate between wakefulness (W), non-rapid-eye-movement sleep (N) and rapid-eye-movement sleep (R). Panel C shows a bidimensional plot of the fraction (density) of 4 s epochs as a function of EMG only; arrows indicate the limits of the boundary region between W and sleep states.

(step 2), which essentially considered isolated 4 s epochs drafted as indeterminate state, NREMS, or REMS, as part of the surrounding wake–sleep state. A Matlab software implementation of the algorithm is available to readers as an Appendix, including full details of each algorithm step and routines for display of EEG spectral analysis results. Readers are encouraged to use this software implementation and contact us for advice or problems with its application.

2.5. Study design

The first step of validation of the SCOPRISM algorithm consisted of three different approaches to evaluate its performance in correctly discriminating among W, NREMS and REMS: (a) computation of sensitivity, specificity, and accuracy; (b) analysis of Bland–Altman plots; and (c) analysis of Receiver Operating Characteristic (ROC) curves. In all these analyses, the correct wake–sleep state discrimination was assumed to be that provided by manual sleep scoring. For each wake–sleep state, the epochs correctly or incorrectly assigned to that state by automatic scoring were defined as true positives (TP) or false positives (FP), respectively, whereas the epochs correctly or incorrectly assigned to a different state by automatic scoring were defined as true negatives (TN) or false negatives (FN), respectively. These parameters allowed us to compute sensitivity (TP/(TP + FN)), specificity (TN/(TN + FP)), and accuracy ((TP + TN)/all epochs) of the SCOPRISM algorithm in each wake–sleep state. The same parameters were also computed between human scorers with the senior scorer (C.B.) considered as the gold standard. The Bland–Altman analysis is usually employed to compare different diagnostic methods by plotting their mean score vs. their difference scores. We employed the Bland–Altman analysis to compare SCOPRISM with manual sleep scoring by plotting the mean and different scores of the time spent in each wake–sleep state computed with the two scoring methods. The ROC curves show how changes in a given parameter modify sensitivity and specificity of a diagnostic method. We employed ROC curves to assess whether the algorithm performance in detecting REMS, which is usually a point of weakness of sleep-scoring algorithms (Rytkonen et al., 2011; Sunagawa et al., 2013), was modified by changing the width of the cutoff region of EEG θ/δ ratio centered around unity (cf. paragraph on automatic sleep scoring above). In particular, we compared the results obtained with the following limits of the cutoff region of EEG θ/δ ratio: 0.95–1.05 (±5%), 0.85–1.15 (±15%), 0.75–1.25 (±25%, default value), and 0.65–1.35 (±35%).

The second step of algorithm validation consisted of comparing estimates of sleep structure (i.e. time spent in each state, sleep episode number and duration, REMS latency, number of CLE) and sleep-dependent cardiovascular variables (i.e., mean values of BP and heart rate in each wake–sleep state) obtained by using manual scoring as previously published (Bastianini et al., 2011; Lo Martire et al., 2012; Silvani et al., 2009) with those we would have obtained by relying only on the automatic sleep-scoring algorithm.

The third step of algorithm validation was to estimate the robustness of SCOPRISM by applying the algorithm to mouse and rat data recorded and scored by 3 independent labs. Since the data from the T.E.S. lab were scored in 10 s epochs, we split each scored epoch into two 5 s epochs with the same wake–sleep state assignment and adapted the SCOPRISM routine for the same epoch-length. Similarly, epoch resolution of manual rat scoring was converted from 1 to 4 seconds. For this purpose, a numeric value was assigned to each manually-scored behavioral state (i.e. 1 = W, 2 = NREMS, 3 = REMS) and median values over consecutive 4 s of recordings were calculated and used for the comparison with SCOPRISM performances. When these median values resulted to be different from any of the 4 values from which they were calculated (i.e. not an integer or a non-scored state), the corresponding 4 s epoch was classified as "undetermined". For each dataset, SCOPRISM performance was evaluated using: (a) Bland–Altman plots; (b) calculation of overall (arithmetic mean between wake–sleep states) accuracy, specificity and sensitivity; and (c) calculation of the difference in the percentage of time spent in each wake–sleep state assigned by manual and automatic scoring method.

2.6. Statistical analysis

T-tests were performed comparing the percentage of recording time spent in W, NREMS and REMS according to SCOPRISM and the corresponding values according to the manual scoring method. Estimates of sleep structure and sleep-dependent cardiovascular variables obtained by relying only on automatic or manual sleep scoring were also compared between groups of mutant mice and their respective controls by applying *t*-tests. A 2-way ANOVA analysis with scoring method (2 levels: SCOPRISM vs. manual scoring) and mouse strain (2 levels: cases vs. controls) was performed to test whether absolute values of the estimates of sleep structure and sleep-dependent cardiovascular variables depended on the scoring method by which they were obtained. A 2-way ANOVA analysis with scoring method (2 levels: SCOPRISM vs. manual scoring) and behavior (3 levels: W, NREMS, REMS) was applied on data of each lab to compare SCOPRISM performance in correctly discriminating the wake–sleep states across laboratories. All results are shown as mean ± SEM with significance at $P < 0.05$.

3. Results

3.1. Identification of outliers

The sensitivity of REMS classification is usually the most critical parameter for automatic sleep-scoring algorithms (Rytkonen et al., 2011; Sunagawa et al., 2013). Accordingly, we found that in 6 mice (7% of cases: 1 TG, 1 KO and 4 WT with hybrid genetic background) SCOPRISM had REMS sensitivity that was more than 2 standard deviations lower than the population mean (Fig. S1, panel A). We then investigated whether in these outlier mice, the 3D plots displaying the fraction of 4 s epochs with a given combination of EEG θ/δ ratio and EMG rms (cf. Fig. 1) had unusual features. In particular, in each of these 6 mice, the ridge positioned at low EMG rms values and extending toward high values of the EEG θ/δ ratio was unusually short (Fig. S1, panel B). This reflected unusually abundant slow EEG activity during REMS (Fig. S1, panel C). Accordingly, in each of these 6 outlier mice, <10% of automatically-scored REMS epochs had an EEG θ/δ ratio >2.5. This was also the case in 2 additional WT mice whose REMS sensitivity, although low (70.1 and 49.3, respectively), differed from the population mean by less than 2 standard deviations (cf. orange dots in Fig. S1, panel A). Based on these results, we elected conservatively to employ this criterion (namely, <10% of automatically-scored REMS epochs with an EEG θ/δ ratio >2.5) as an exclusion criterion for the correct application of the automatic sleep-scoring algorithm. Based on this criterion, therefore, we excluded 8 mice from the following analyses.

In Bland–Altman plots, 2 mice were extreme outliers (i.e., with values differing more than 2 standard deviations from the population mean) regarding the time spent in W and NREMS (cf. red dots in Fig. 2, panels A and B, respectively). We again investigated whether the 3D plots of these mice had unusual features that potentially permitted their identification. We found that these 2 mice were the only ones in which the 3D plot showed 3 peaks instead of 2 (Fig. S2, panel A). This anomaly was caused by inappropriately high EMG signal amplification (Fig. S2, panel B), which led to an erroneous assignment of the EMG rms boundary zone region (cf. red arrow in Fig. S2, panel C), and eventually to a faulty W-NREMS distinction.

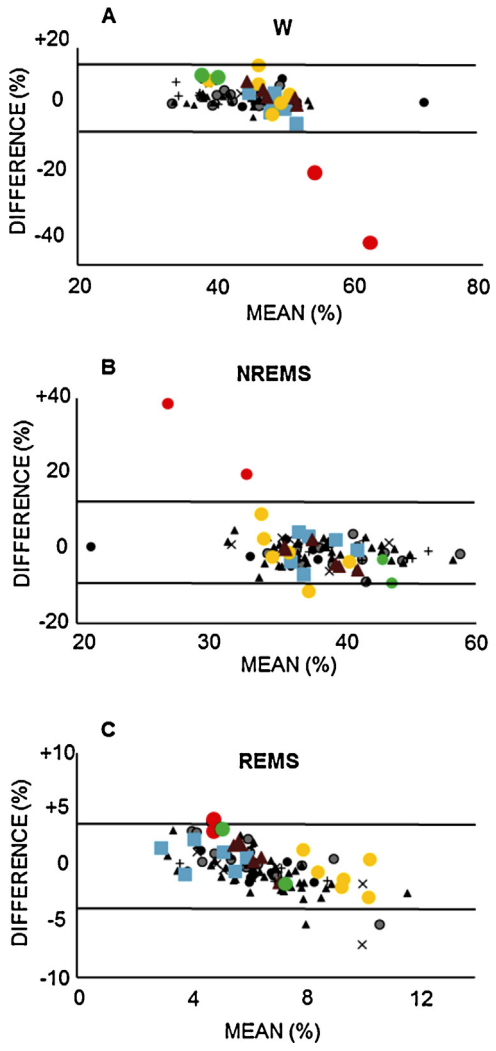

Fig. 2. Bland–Altman plots of 89 mice during wakefulness (W), non-rapid-eye-movement sleep (NREMS) and rapid-eye-movement sleep (REMS). The plots show the mean (*x* axis) vs. the difference (*y* axis) of the percentage of recording time assigned by manual and automatic scoring to each wake–sleep state. Legend: black circles, TG mice; gray circles, HTG mice; X-shaped symbols, KO mice; crosses, ob/ob mice; gray triangles, WT mice; light blue rectangles, WT mice from T.E.S lab; brown triangles, WT mice from M.K. lab; orange circles, rats from R.A. lab. Dots relative to the two mice with 3 peaks in the tridimensional plot are shown before (red) and after (green) correction of the boundary zone between W and sleep in the scoring algorithm (cf. Section 3.1). Horizontal lines delimit the 95% confidence interval of the distribution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We thus repeated the automatic sleep scoring on these 2 mice after imposing a shift of the EMG rms boundary zone region toward the peak with highest EMG rms values (cf. green arrow in Fig. S2, panel C). This change resolved the outlier condition of these 2 mice (cf. green dots in Fig. 2, panels A and B), and was thus applied before including these mice in the subsequent analyses.

3.2. Internal validation of SCOPRISM algorithm

Table 1 shows values of accuracy, specificity and sensitivity between human scorers and those between SCOPRISM and human scorers in detecting W, NREMS and REMS epochs. All these values were very high and comparable between manual and automatic scoring. Bland–Altman plots (Fig. 2) did not highlight any difference in the algorithm performance related to the mouse strain. There were very small yet significant effects of the scoring method on the percentage of recording time spent in W (−1.8% with automatic scoring) and NREMS (+1.0% with automatic scoring). ROC curve analysis demonstrated that the smaller the width of the boundary region based on EEG θ/δ ratio, the higher the REMS sensitivity (Fig. S3, panel A). Over the 4 different boundary region widths that we evaluated, REMS sensitivity increased at most by 3.2%, but this came at the expense of REMS specificity (−2.3%).

Table 1
Accuracy, specificity and sensitivity of manual and automatic sleep scoring.

		W	NREMS	REMS
AUTOMATIC (%)	ACC	97.6 ± 0.2	99.4 ± 0.1	95.4 ± 0.5
	SPEC	95.3 ± 0.3	94.8 ± 0.5	95.7 ± 0.4
	SENS	97.9 ± 0.1	98.6 ± 0.2	86.2 ± 1.7
MANUAL (%)	ACC	98.0 ± 0.3	97.6 ± 0.3	99.6 ± 0.1
	SPEC	97.2 ± 0.4	98.5 ± 0.3	99.6 ± 0.1
	SENS	99.3 ± 0.2	95.6 ± 0.6	87.5 ± 2.9

W, Wakefulness; NREMS, non-rapid-eye-movement sleep; REMS, rapid-eye-movement sleep; ACC, accuracy; SPEC, specificity; SENS, sensitivity; AUTOMATIC, automatic sleep scoring; MANUAL, manual sleep scoring.

3.3. The application of SCOPRISM algorithm to investigate changes in sleep structure and sleep-dependent cardiovascular control in mouse models of human disease

There were no significant effects of the scoring method either on estimated duration of wake–sleep episodes or estimated cardiovascular values in each wake–sleep state (Tables S1–S6 and Fig. S4). The automatic and manual scoring yielded small yet significant differences in estimates of REMS latency (population means, 498 vs. 447 s, respectively), number of W episodes per day (population means, 168 vs. 184 episodes/day, respectively) and number of NREMS episodes per day (population means, 466 vs. 482 episodes/day, respectively). By relying only on automatic scoring, we were able to replicate the main findings of our previously published studies (Bastianini et al., 2011; Lo Martire et al., 2012; Silvani et al., 2009) in terms of sleep structure and of sleep-dependent cardiovascular variables between mutant mice and their WT controls (Tables S1–S6). In particular, the automatic sleep-scoring algorithm recognized all main features of the narcolepsy phenotype in TG and KO mice (Tables S1, S2, S4, and S5) including CLE (Table S3).

3.4. SCOPRISM validation on independent mouse and rat datasets.

All the mice recorded in the M.K. lab and all the rats recorded in the R.A. lab satisfied the SCOPRISM inclusion criteria (cf. Section 3.1) while we excluded 2 of the 8 mice recorded in the T.E.S lab because they had <10% of automatically-scored REMS epochs with EEG θ/δ ratio > 2.5 (cf. Section 3.1). Initially, we applied SCOPRISM on mouse and rat data from outside the PRISM lab without changing any algorithm parameters but we obtained suboptimal results on rat (data not shown). Visual inspection of raw tracings showed higher theta power in the rat EEG compared to the mouse EEG. Accordingly, using a ROC curve analysis we established that in rats, the best sleep state discrimination was reached by centering the NREMS-REMS boundary region at $\theta/\delta = 2$ instead of using $\theta/\delta = 1$ (Fig. S3, panel B). With this correction, we determined that in the Bland–Altman plots (Fig. 2) all the dots relative to mice and rats (except in 1 case in panel B) studied in independent laboratories fell within the 95% confidence interval calculated from PRISM lab data. SCOPRISM identified the same total amount of each behavioral state than those manually scored by independent mouse labs with the exception of a slightly lower amount of W (Table S7) scored

on mice recorded in the M.K. lab and rats recorded in the R.A. lab. The overall accuracy of SCOPRISM, its specificity, and its sensitivity obtained in the PRISM lab (97%, 94% and 95%, respectively) were very similar to those calculated on mouse data recorded in the M.K. lab (96%, 97% and 97%, respectively) and only slightly higher than those calculated on mouse data recorded in the T.E.S. lab (92%, 91% and 88%, respectively) and rat data (89%, 87%, and 95%, respectively) recorded in the R.A. lab.

4. Discussion

We developed and cross-validated SCOPRISM, a new, robust and reliable algorithm for automatic scoring of mouse sleep.

Several automatic algorithms have been proposed in the last few years (Brankack et al., 2010; Rytkonen et al., 2011; Sunagawa et al., 2013; Veasey et al., 2000) but issues such as complexity, subjectivity and robustness to different mouse mutants and datasets still represent a barrier to their widespread adoption. The present algorithm validation addressed all these points. In fact, we developed a simple, open-source and fully-automated algorithm that can be implemented in programming environments such as Matlab (Appendix). SCOPRISM attempts to improve on a previous algorithm developed by Veasey et al. (2000) by avoiding the risk of subjectivity in determining wake–sleep cutoffs. SCOPRISM performed similarly to the algorithm proposed by Veasey et al. (2000) in terms of W and NREMS accuracy (−1.8% and +2.2%, respectively), but performed better in terms of REMS (+5.7%) accuracy. Moreover, SCOPRISM performed better in terms of overall accuracy and sensitivity and slightly worse in terms of specificity (97.4%, 94.2% and 95.3%, respectively) compared to more recent algorithms by Rytkonen et al. (2011) (94.6%, 92.5% and 97.5%, respectively) or by Sunagawa et al. (2013) (93.6%, 93.1% and 97.2%, respectively). We systematically assessed whether sleep structure and sleep-dependent cardiovascular variables estimated using automatic sleep scoring differed from those estimated with manual scoring (Bastianini et al., 2011; Lo Martire et al., 2012; Silvani et al., 2009). With our algorithm, we were able to replicate the main results in terms of sleep structure and sleep-dependent cardiovascular variables (Tables S1–S6) that we previously found with manual scoring (Bastianini et al., 2011; Lo Martire et al., 2012; Silvani et al., 2009). Remarkably, SCOPRISM performed well in recognizing peculiarities of the narcolepsy phenotypes such as W fragmentation, reduced REMS latency, and CLE (Table S3). The robustness of our algorithm is supported by its validation using a large dataset of 89 mice of different strains, including, for the first time, validation on 3 different genetically-modified strains of mice, 2 of which displayed overt sleep disturbances (Chemelli et al., 1999; Hara et al., 2001). We did not find any SCOPRISM performance difference related to mouse strains.

By testing SCOPRISM on a larger population of mice than those studied in the past for this purpose (Brankack et al., 2010; Rytkonen et al., 2011; Sunagawa et al., 2013), we had the chance to evaluate relatively rare alterations in raw EEG and EMG signal quality that may affect algorithm performance. In fact, we found 2 types of such alterations (Figs. S1 and S2), and identified their respective fingerprint on the 3D plot, which shows the fraction of 4 s epochs

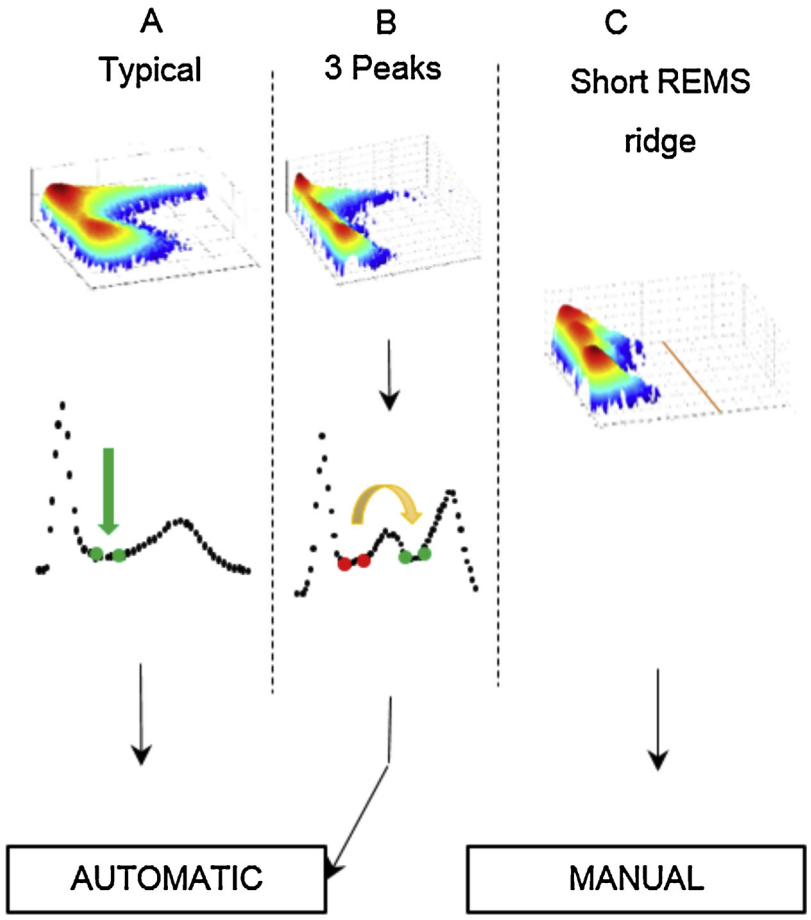

Fig. 3. Flow chart for the correct application of the automatic sleep-scoring algorithm. Three different scenarios are reported: typical distribution (A), distribution with 3 peaks (B), distribution with short ridge corresponding to rapid-eye-movement sleep (REMS, C). In cases A and B, if the boundary zone between wakefulness and non-rapid-eye-movement sleep is correctly set (green dots), automatic sleep scoring can be safely applied. In case C, we suggest manually scoring sleep states. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

characterized by given combinations of EEG θ/δ ratio and EMG rms. This led us to propose a visual flow chart (Fig. 3) for the correct application of the current algorithm based on objective evaluation of the 3D plots. In particular, our results suggested that the 3D graph of each given mouse may conform to 3 possible cases: (a) it may be typical and similar to that displayed in Fig. 1; (b) it may show 3 distribution peaks instead of 2 (Fig. S2, panels A and C); (c) it may show a short REMS ridge (Fig. S1, panel B). In the case of (a), our analysis suggests that performance of our automatic sleep-scoring algorithm can be considered reliable, and no corrections are needed. In case (b), it is necessary to ensure manually that the EMG rms boundary region is correctly set next to the peak with the highest EMG rms values. If so, we suggest that the performance of our automatic sleep-scoring algorithm be considered reliable. In case (c), it is necessary to determine the percentage of automatically scored REMS epochs with an EEG θ/δ ratio >2.5 (this computation may be performed automatically by the software routine, see Appendix). If this percentage is lower than 10%, our results suggest that the quality of the raw EEG signal is not suited for automatic sleep scoring, and, thus, that these animals should be excluded from the experiment unless it has been proved that the scoring method, manual or automatic, does not affect the variables of interest. In this case, manual scoring of these specific animals can be performed and merged with automatic scoring of other animals included in the same experiment. This flow chart is quick and easy to follow, and most importantly, it allows experimenters to evaluate the reliability of the automatic sleep scoring before knowing the results of the sleep analysis, thus preventing any bias. This peculiarity is entirely new for an automatic sleep-scoring algorithm. It is worth noting that the typical 3D plot (Fig. 1) also provides interesting insights into EEG and EMG features of the different wake–sleep states in mice. In particular, W and NREMS can be easily distinguished on the basis of their EMG properties in mice, even though W is the behavioral state with the widest epoch distribution in the 3D plot, reflecting its inherent variety. On the other hand, no clear boundary is present in the 3D plot between the NREMS peak and the REMS ridge, indicating that NREMS and REMS represent a continuum in terms of EEG (power in the θ and δ frequency ranges) and EMG properties in mice.

Finally, since in our previous studies we manually scored wake–sleep states based on data pre-scored by the present algorithm, we performed a cross-lab validation of SCOPRISM to exclude any internal bias in our manual scoring. To the best of our knowledge, SCOPRISM is the first automatic sleep-scoring algorithm tested on mouse and rat datasets recorded and analyzed by 3 independent laboratories. SCOPRISM performed well when applied to both mouse data of independent labs and rat data (Fig. 2 and Table S7). The SCOPRISM overall values of accuracy, specificity and sensitivity were similar in our lab and in the M.K. lab and slightly lower in the T.E.S. lab. This discrepancy may be explained, at least in part, by the different epoch length of the manual scoring (4 s in our lab and in the M.K. lab, 10 s in the T.E.S. lab) that required a computational adjustment before the SCOPRISM application (i.e. we split each 10 s scored epoch into two 5 s epochs, cf. Section 2.5). The good performance of SCOPRISM on T.E.S. lab data is further attested by the absence of statistical differences in the percentage of time spent in each wake–sleep state (mean values in Table S7 and Fig. 2) between the automatic and the manual scoring. The application of SCOPRISM on rat data required extra analysis to define the best cut-off region between NREMS and REMS (Fig. S3, panel B). However, after this upgrade, SCOPRISM performances on rat data reached values similar to those calculated on mice (cf. Section 3.4).

Several limitations of SCOPRISM have to be acknowledged. Firstly, SCOPRISM relies heavily on the EEG θ/δ ratio, which may change depending on the position of the recording electrodes and the use of differential or bipolar EEG montages. This may partly explain the variability of SCOPRISM performance within and between laboratories and species. Moreover, SCOPRISM was developed for off-line analyses. However, our algorithm is suitable for on-line applications such as real-time REMS deprivation. Indeed, since the first step of SCOPRISM operation is entirely based on EEG and EMG properties of each 4-s epoch (cf. Section 2.4) it is possible to use the discriminating criteria, calculated for each animal during a baseline recording, to implement specific on-line applications. SCOPRISM also shares intrinsic limitations of all automatic sleep scoring algorithms. In particular, automatic sleep scoring does not relieve investigators from careful inspection of raw EEG/EMG traces, which may be crucial to ensure precise characterization of wake–sleep behavior. Moreover, rigidly fixed algorithms for automatic sleep scoring may be inadequate for particular sleep analyses. For example, inclusion of epochs containing two sleep–wake stages or movement artifacts may bias results of the EEG spectral analysis, and scoring short awakenings during NREMS as wakefulness may apparently enhance sleep fragmentation in analyses of sleep episode number and duration. In these respects, however, SCOPRISM may take advantage from its being open source, its specific consideration of indeterminate states, and its criteria for inclusion of isolated epochs drafted as indeterminate state into the surrounding wake–sleep state (cf. Section 2.4).

4.1. Conclusions

We developed and validated SCOPRISM, a new, open-source and robust algorithm for sleep scoring in mice. We successfully validated this algorithm on mouse and rat data independently recorded and analyzed by other 3 labs. Finally, we suggest the use of SCOPRISM to accelerate preclinical studies on sleep research.

Author contributions

Conceived and designed the study: SB, AS, GZ conceived and designed the study. SB, CB, FDV, VLM, CA, MG performed manual sleep scoring. AS designed and implemented the automatic sleep-scoring algorithm. SB performed data analysis. SB wrote the paper. All authors reviewed the manuscript.

Acknowledgements

This work has been funded by the University of Bologna (RFO 08-11; FFBO120705), Ministry of Instruction, University and Research, Italy (PRIN 2008, prot. 2008FY7K9S)

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jneumeth.2014.07.018.

References

Bastianini S, Silvani A, Berteotti C, Elghozi JL, Franzini C, Lenzi P, et al. Sleep related changes in blood pressure in hypocretin-deficient narcoleptic mice. Sleep 2011;34:213–8.

Brankack J, Kukushka VI, Vyssotski AL, Draguhn A. EEG gamma frequency and sleep–wake scoring in mice: comparing two types of supervised classifiers. Brain Res 2010;1322:59–71.

Cerri M, Mastrotto M, Tupone D, Martelli D, Luppi M, Perez E, et al. The inhibition of neurons in the central nervous pathways for thermoregulatory cold defense induces a suspended animation state in the rat. J Neurosci Off J Soc Neurosci 2013;33:2984–93.

Chemelli RM, Willie JT, Sinton CM, Elmquist JK, Scammell T, Lee C, et al. Narcolepsy in orexin knockout mice: molecular genetics of sleep regulation. Cell 1999;98:437–51.

El Helou J, Belanger-Nelson E, Freyburger M, Dorsaz S, Curie T, La Spada F, et al. Neuroligin-1 links neuronal activity to sleep–wake regulation. Proc Natl Acad Sci U S A 2013;110:9974–9.

Fenzl T, Romanowski CP, Flachskamm C, Honsberg K, Boll E, Hoehne A, et al. Fully automated sleep deprivation in mice as a tool in sleep research. J Neurosci Methods 2007;166:229–35.

Gondard E, Anaclet C, Akaoka H, Guo RX, Zhang M, Buda C, et al. Enhanced histaminergic neurotransmission and sleep–wake alterations, a study in histamine H3-receptor knock-out mice. Neuropsychopharmacology 2013;38:1015–31.

Hara J, Beuckmann CT, Nambu T, Willie JT, Chemelli RM, Sinton CM, et al. Genetic ablation of orexin neurons in mice results in narcolepsy, hypophagia, and obesity. Neuron 2001;30:345–54.

Hara J, Yanagisawa M, Sakurai T. Difference in obesity phenotype between orexin-knockout mice and orexin neuron-deficient mice with same genetic background and environmental conditions. Neurosci Lett 2005;380:239–42.

Kantor S, Mochizuki T, Lops SN, Ko B, Clain E, Clark E, et al. Orexin gene therapy restores the timing and maintenance of wakefulness in narcoleptic mice. Sleep 2013;36:1129–38.

Kimura M, Muller-Preuss P, Lu A, Wiesner E, Flachskamm C, Wurst W, et al. Conditional corticotropin-releasing hormone overexpression in the mouse forebrain enhances rapid eye movement sleep. Mol Psychiatry 2010;15:154–65.

Lo Martire V, Silvani A, Bastianini S, Berteotti C, Zoccoli G. Effects of ambient temperature on sleep and cardiovascular regulation in mice: the role of hypocretin/orexin neurons. PLoS ONE 2012;7:e47032.

Rytkonen KM, Zitting J, Porkka-Heiskanen T. Automated sleep scoring in rats and mice using the naive Bayes classifier. J Neurosci Methods 2011;202:60–4.

Scammell TE, Willie JT, Guilleminault C, Siegel JM, International Working Group on Rodent Models of N. A consensus definition of cataplexy in mouse models of narcolepsy. Sleep 2009;32:111–6.

Silvani A, Bastianini S, Berteotti C, Franzini C, Lenzi P, Lo Martire V, et al. Sleep modulates hypertension in leptin-deficient obese mice. Hypertension 2009;53:251–5.

Sunagawa GA, Sei H, Shimba S, Urade Y, Ueda HR. FASTER: an unsupervised fully automated sleep staging method for mice. Genes Cells 2013;18:502–18.

Veasey SC, Valladares O, Fenik P, Kapfhamer D, Sanford L, Benington J, et al. An automated system for recording and analysis of sleep in mice. Sleep 2000;23:1025–40.